Esercizi - Analisi Predittiva

Modelli lineari semplici e multipli

Appunti per la Soluzione ad esercizi scelti

Esercizio 1

I dati in `grain.dat` sono stati raccolti nel 2007 in uno studio sula relazione tra la resa in termini di alcool nel processo di distillazione e l'azoto contenuto nel grano distillato. I dati sono stati raccolti in quattro diverse aree del Regno Unito. Il dataset ha tre colonne: `nitrogen` è la percentuale di azoto (per kilogrammo), `alcohol` è la resa in alcool in Litri per Tonnellata, e`location` indica il luogo in cui è stato coltivato il grano. [Il dataset è stato reso disponibile da Julian Faraway.]

La relazione tra la resa in termini di alcool e l'azoto contenuto nel grano può essere indagata con il seguente modello lineare:

$$\text{alcohol}_i = \alpha + \beta \ \text{nitrogen}_i + \epsilon_i \tag{1}$$

1. Si produca un grafico dei dati. La relazione tra le variabili in esame appare lineare?
2. Si dia una stima puntuale per α e β.
3. Si dia una stima intervallare ad un livello di confidenza di 99% per α e β.
4. Quali sono le assunzioni necessarie per poter aver stime puntuali per i valori α e β? Quali sono le assunzioni necessarie per poter ottenere delle stime intervallari per α e β?
5. Si aggiunga la retta delle relazione stimata tra `alcohol` e `nitrogen` al grafico ottenuto al punto 1.
6. Il dataset contiene la variabile `location`. Si scriva in forma estesa il modello che R stima quando si usa la funzione `lm(alcohol location, data = grain)`.
7. É valida l'affermazione che la variabile `location` spiega una buona parte della variabilità della variabile `alcohol`?
8. Se si aggiunge la variabile `location` al modello in eq. (1) in cui solo `nitrogen` era presente nel modello, l'aggiunta di `location` risulta significativa? Come si può misurare l'evidenza contro la non-inclusione di `location` nel modello?
9. Si produca un grafico della relazione tra `location` e `nitrogen` - cosa si può notare?
10. Come si spiega la differenza dei p-value per `location` nei modelli stimati al punto 6 e al punto 8?
11. Usando il modello specificato in eq. (1): si predica il valore medio della resa di alcool per del grano contenente il 1.9% e il 2.7% di azoto per kilogrammo.

12. Si stimino gli intervalli di confidenza al 95% per i valori medi della resa di alcool stimati al punto 11. Quale è l'ampiezza di questi intervalli: si spieghi la differenza nell'ampiezza.

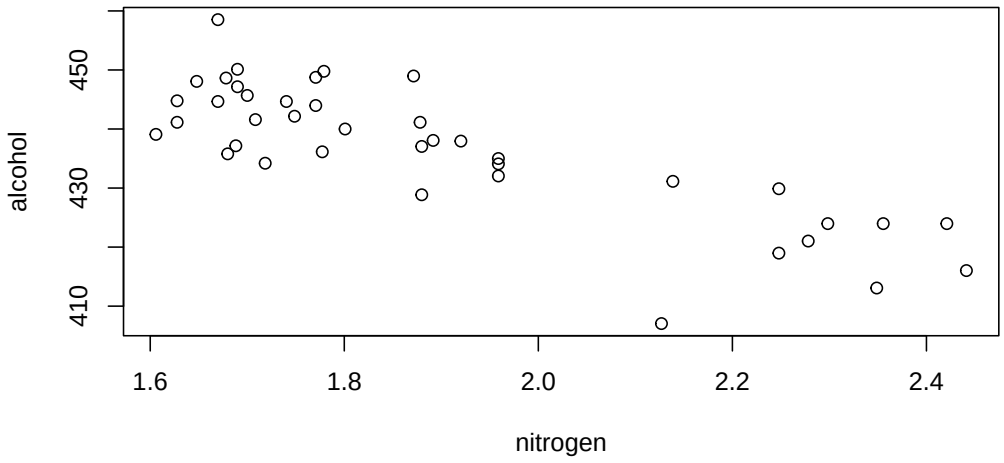
13. Usando il modello specificato in eq. (1): si predica il valore effettivo della resa di alcool per del grano contenente il 1.9% e il 2.7% di azoto per kilogrammo. Si dia una anche una valutazione degli intervalli predittivi al 95% per questi valori.

Soluzione 1

```
grains <- read.table("grains.dat", header = TRUE)
```

1. La relazione tra `alcohol` e `nitrogen` appare abbastanza lineare.

   ```
   plot(grains[,c("nitrogen","alcohol")])
   ```




   ```
   fit <- lm(alcohol ~ nitrogen, data = grains)
   coef(fit)
   ```

   ```
   (Intercept)    nitrogen
     506.87058   -37.03356
   ```

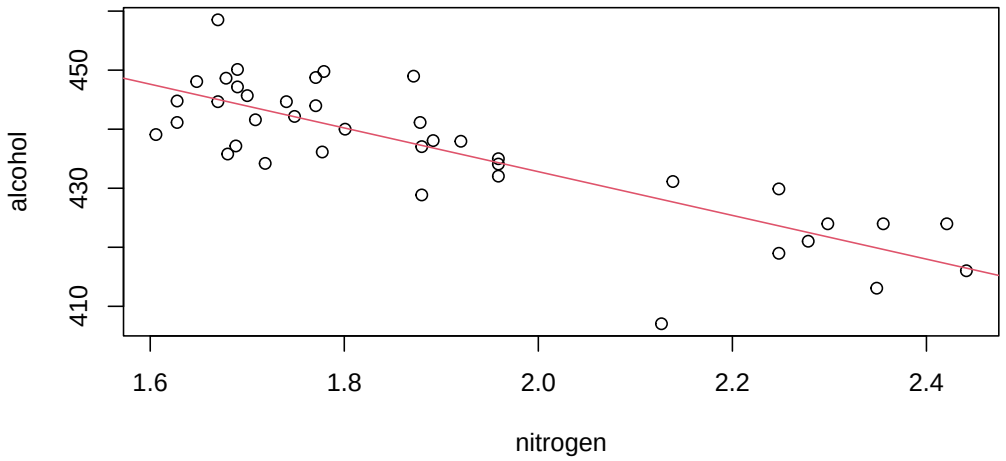
2. $\hat{\alpha} = 506.87$ e $\hat{\beta} = 506.87$

3. ```
   confint(fit)
   ```

   ```
                    2.5 %    97.5 %
   (Intercept) 491.00147 522.73969
   nitrogen    -45.32595 -28.74117
   ```
   ```

4. Per poter ottenere stime puntuali di α e β assumiamo che i dati $(y_1, \ldots, y_n)$ siano indipendenti tra loro, identicamente distribuiti con varianza costante e valore atteso che cambia approssimativamente linearmente con X. Per poter costruire delle stime intervallare dobbiamo anche assumere che i dati siano distribuiti secondo una normale.

5. ```
   plot(grains[,c("nitrogen","alcohol")])
   abline(fit, col = 2)
   ```



6. $$\text{alcohol}_i = \alpha + \beta \ \text{location}_i + \epsilon_i \tag{2}$$

   con ϵ_i iid $\epsilon_i \sim N(0, \sigma^2)$

7. ```
   fit2 <- lm(alcohol~location, data = grains)
   summary(fit2)$r.square
   ```

   ```
   [1] 0.7281804
   ```

   Sì, la variabile `location` spiega una buona parte della variabilità della variabile `alcohol`, circa il 73%.

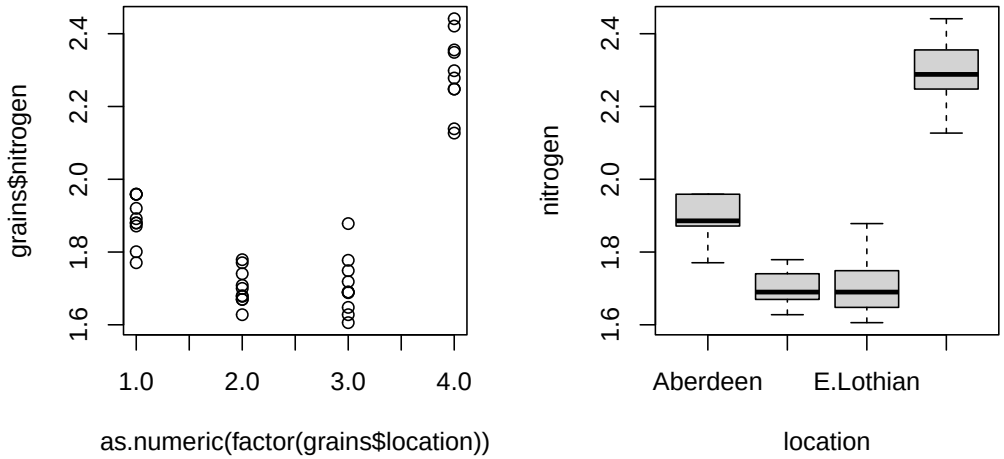
8. ```
   fit3 <- lm(alcohol~nitrogen+location, data = grains)
   anova(fit2, fit3)
   ```

   ```
   Analysis of Variance Table

   Model 1: alcohol ~ location
   Model 2: alcohol ~ nitrogen + location
     Res.Df    RSS Df Sum of Sq      F Pr(>F)
   1     36 1367.5
   2     35 1338.2  1    29.305 0.7664 0.3873
   ```
   ```

Se aggiungiamo location al modello in equazione (1), questa aggiunta non risulta significativa giacché la RSS diminuisce di poco (relativamente alla variabilità del sistema): possiamo misurare l'evidenza contro l'inclusione di `location` tramite RSS e il test F (anova table).

9. ```
   par(mfrow = c(1,2)) #  due opzioni
   plot(as.numeric(factor(grains$location)), grains$nitrogen)
   boxplot(nitrogen~location, data = grains)
   ```



Notiamo che `location` e `nitrogen` hanno una relazione tra loro: le due variabili sono parzialmente co-lineari.

10. Dato che `location` è co-lineare con `nitrogen` sebbene la variabile sia significativa quando inserita nel modello come unico predittore, essa diventa non significativa quando viene inserita una combinazione con una variabile con cui è legata.

11. ```
    predict(fit, newdata = data.frame(nitrogen = c(1.9,2.7)))
    ```

    ```
           1        2
    436.5068 406.8800
    ```

12. ```
    ci <- predict(fit, newdata = data.frame(nitrogen = c(1.9,2.7)), interval = "confidence")
    ci; ci[,3]-ci[,2]
    ```

    ```
           fit      lwr      upr
    1 436.5068 434.4319 438.5818
    2 406.8800 399.9075 413.8524

            1         2
     4.149899 13.944844
    ```
    ```

Nel campione osservato non si hanno osservazioni per cui sono stati usati più di 2.44% azoto per kilogrammo: la stima per la seconda osservazione è molto più incerta perché basata sull'estrapolazione del modello stimato. oltre i limiti della variabile risposta per cui è stato stimato. Al contrario invece il valore di 1.9% è molto vicino al valore medio dell'azoto misurato per il campione: per questi valori la stima ha la minor incertezza possibile.

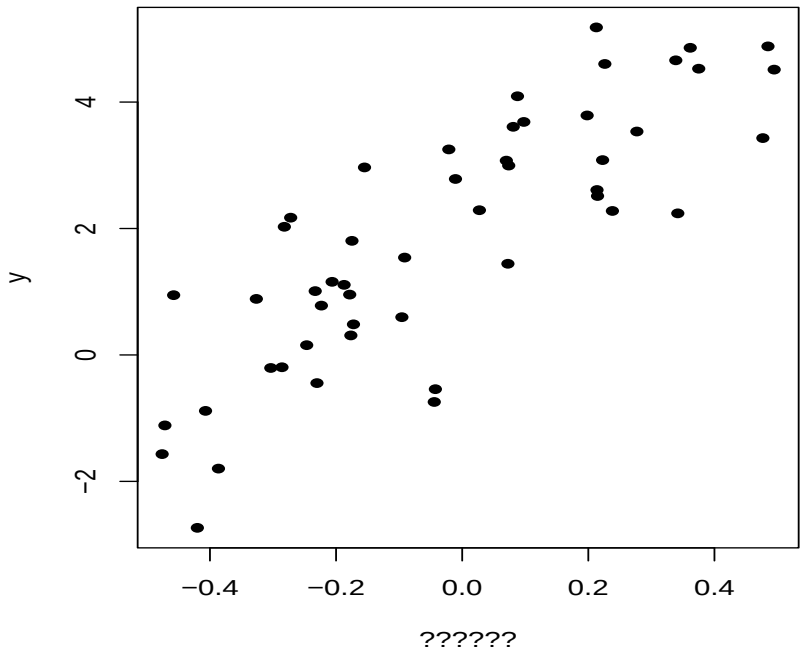
13. ```
    pi <- predict(fit, newdata = data.frame(nitrogen = c(1.9,2.7)), interval = "predict")
    pi; pi[,3]-ci[,2]
    ```

    ```
           fit      lwr      upr
    1 436.5068 423.2214 449.7922
    2 406.8800 392.0203 421.7397

           1        2
    15.36034 21.83213
    ```

    Anche negli intervalli di predizione notiamo che quando il modello vine usato per predire valori al di fuori dell'intervallo osservato nei dati originali questa predizione sarà molto più incerta.

Esercizio 3



1. Il grafico qui sopra mostra la relazione tra la variabile X e Y di interesse. Qui sotto vengono riportati i `summary` di due modelli stimati: uno usando la X mostrata in figura e uno usando un'altra variabile. Si identifichi il summary che corrisponde alla relazione mostrata in figura.

   S1
    ```

```
summary(lm(y ~ x1))

Call:
lm(formula = y ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.49684 -0.68150  0.03744  0.78701  1.88648

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0185     0.1595   12.65  < 2e-16 ***
x1            5.9989     0.5858   10.24 1.16e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.122 on 48 degrees of freedom
Multiple R-squared:  0.686,        Adjusted R-squared:  0.6795
F-statistic: 104.9 on 1 and 48 DF,  p-value: 1.157e-13
```

S2

```
summary(lm(y ~ x2))

Call:
lm(formula = y ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9848 -1.6183  0.4791  1.2761  3.7511

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.50848    1.08136   3.245  0.00215 **
x2          -0.03306    0.02086  -1.585  0.11960
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.952 on 48 degrees of freedom
Multiple R-squared:  0.04972,        Adjusted R-squared:  0.02992
F-statistic: 2.511 on 1 and 48 DF,  p-value: 0.1196
```

2. Come si può interpretare il seguente output di R?

```
anova(lm(y ~ x1), lm(y ~ x1+x2))

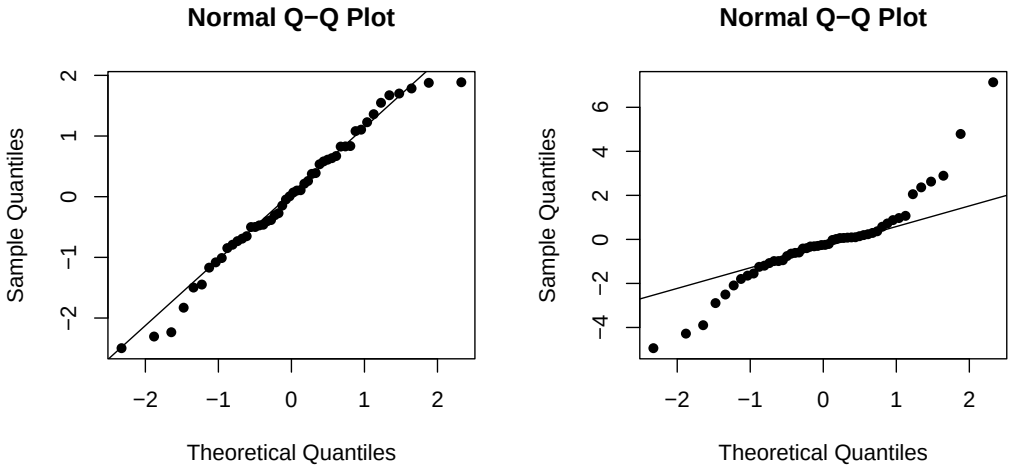
Analysis of Variance Table
```

```
Model 1: y ~ x1
Model 2: y ~ x1 + x2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     48 60.451
2     47 58.730  1    1.7209 1.3772 0.2465
```

3. Qui sotto vengono mostrati tre valori di R^2 e i tre modelli da cui sono stati estratti: si accoppino i modelli e valori di R^2 ad essi corrispondenti:

Modello	R^2
`lm(y ~ x1+x2)`	0.686
`lm(y ~ x1)`	0.0497
`lm(y ~ x2)`	0.695

4. Quale dei grafici quantile-quantile indica un comportamento del campione analizzato più simile alla distribuzione di riferimento? Si spieghi come i grafici quantile-quantile possono essere utilizzati quando si stimano modelli lineari.



Soluzione 3

1. Il summary corrispondente alla figura deve essere **S1**: la relazione tra X e Y è positiva, ci sia aspetta che $\hat{\beta} > 0$
2. Dall'ouput possiamo dedurre che l'inclusione si una variabile $x2$ non è significativa
3. Il modello più complesso avrà sicuramente il valore di R^2 più alto, inoltre sappiamo che il modello ha un valore di R^2 di .686 (e dal grafico sappiamo che la relazione tra X e Y è abbastanza forte). Si ha quindi

Modello	R^2
`lm(y ~ x1+x2)`	0.695
`lm(y ~ x1)`	0.0497
`lm(y ~ x2)`	0.0497

4. Il grafico sulla sinistra indica un comportamento del campione simile alla disitrbuzione di riferimento (in questo caso, la normale). I punti infatti si allineano sulla bisettrice, indicando che i valori empirici dei quantili del campione corrispondo ai valori che ci si potrebbe aspettare teoricamente estraendo un campione di dimensione n dalla disitrbuzione di riferimento. Il grafico sulla destra invece mostra che il campione ha delle code più pesanti della disitrbuzione di riferimento.

Esercizio 4

Si prenda in esame il dataset `prostate` dal pacchetto R `faraway`. Si desidera stimare la relazione tra la un certo antigene (descritto dalla variabile `lpsa`) e altre variabili contenute nel dataset.

1. Si prenda in considerazione un modello lineare multiplo in cui tutte le variabili presenti nel dataset sono usate come predittori (Modello 1). Si stimi il modello e usando la funzione `summary` (o equivalenti) si trovi il valore della statistica F della significativitá globale del modello. Si interpreti il valore della statistica test.

2. Si trovi la stima puntuale del coefficiente di regressione legato alla variabile `age` dentro al modello 1: che interpretazione si può dare al valore del coefficiente? Si produca un grafico di dispersione (scatter plot) della variabile `age` e la variabile `lpsa`: come si può interpretare il coefficiente di regressione identificato per il modello 1 alla luce del grafico?

3. Si trovi l'intervallo di confidenza al 90% e 99% per il coefficiente di regressione legato alla variabile `age` dentro al modello 1: che interpretazione si può dare ai due intervalli? Cosa si può dedurre da questi intervalli di confidenza sul p-value della variable `age` nel Modello 1?

4. Si stimino intervalli di confidenza per il valore di `lpsa` di due pazienti con le seguenti caratteristiche:

```
nd <- prostate[1,-9] ## to ensure correct names
nd[1,] <- c(1.45, 3.62,65,0.3,0,-0.8,7,15)
nd[2,] <- c(4,6.2,83,2.33,1,2.96,9,100)
rownames(nd) <- c("Patient A", "Patient B")
nd
```

```
          lcavol lweight age lbph svi   lcp gleason pgg45
Patient A   1.45    3.62  65 0.30   0 -0.80       7    15
Patient B   4.00    6.20  83 2.33   1  2.96       9   100
```

   Si commenti l'ampiezza degli intervalli di confidenza.

5. Si stimi ora un nuovo modello, modello 2, in cui solo le variabili predittive con un p-value del test di significatività nel Modello 1 minore di 0.05.

6. Si usi questo nuovo modello per stimare i valori dei pazienti A e B: si commenti sull'ampiezza degli intervalli di confidenza trovati nel Modello 1 e Modello 2.

7. Si testi al significatività del modello 2 contro il modello 1, esplicitando l'ipotesi nulla e alternativa sotto studio.

   Per caricare il dataset nella propria workspace è necessario avere il pacchetto faraway installato - questo si può fare una volta sola con il comando install.packages("faraway"). Successivamente sarà necessario usare il comando data(prostate, package = "faraway").

Soluzione 4

```
data(prostate, package = "faraway")
```

1. ```
   fit_all <- lm(lpsa ~ ., data = prostate)
   summary(fit_all)
   ```

   ```
   Call:
   lm(formula = lpsa ~ ., data = prostate)

   Residuals:
       Min      1Q  Median      3Q     Max
   -1.7331 -0.3713 -0.0170  0.4141  1.6381

   Coefficients:
               Estimate Std. Error t value Pr(>|t|)
   (Intercept)  0.669337   1.296387   0.516  0.60693
   lcavol       0.587022   0.087920   6.677 2.11e-09 ***
   lweight      0.454467   0.170012   2.673  0.00896 **
   age         -0.019637   0.011173  -1.758  0.08229 .
   lbph         0.107054   0.058449   1.832  0.07040 .
   svi          0.766157   0.244309   3.136  0.00233 **
   lcp         -0.105474   0.091013  -1.159  0.24964
   gleason      0.045142   0.157465   0.287  0.77503
   pgg45        0.004525   0.004421   1.024  0.30886
   ---
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   Residual standard error: 0.7084 on 88 degrees of freedom
   Multiple R-squared:  0.6548,        Adjusted R-squared:  0.6234
   F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
   ```

   La statistica test per il test della significativitá globale è piuttosto grande, ad indicare che, per qualunque livello di significativitá usato comunemente, si può rigettare l'ipotesi nulla che tutti i coefficienti di regressione siano pari a 0: il modello cattura una porzione rilevante della variabilitá dei dati.

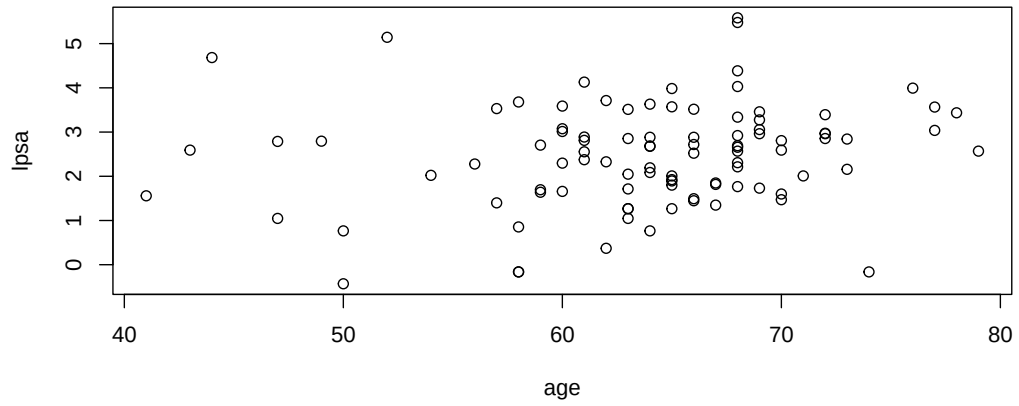
2. La stima per il parametro relativo ad `age` è

   ```
   coef(fit_all)["age"]
   ```

   ```
          age
   -0.01963718
   ```

   sebbene la relazione tra le due variabili sembri positiva

   ```
   plot(prostate[,c("age","lpsa")])
   ```
   ```



Probabilmente siamo in presenza di probelmi di co-linearità multipla tra i predittori.

3.
```
confint(fit_all, "age",level = .90)
```

```
           5 %          95 %
age -0.0382102 -0.001064151
```

```
confint(fit_all, "age",level = .99)
```

```
          0.5 %      99.5 %
age -0.04905337 0.009779023
```

```
coef(summary(fit_all))["age",]
```

```
   Estimate  Std. Error     t value    Pr(>|t|)
-0.01963718  0.01117272 -1.75759949  0.08229321
```

L'intervallo di confidenza al 99% per il parametro contiene 0, mentre l'intervallo di confidenza al 90% non contiene 0: ne deduciamo che il p-value per la verifica di ipotesi $H_0 : \beta_{age} = 0 \; VS \; H_1 : \beta_{age} \neq 0$ è più grande di 0.01 ma più piccolo di 0.1 (il valore è infatti 0.08 circa).

4. Deriviamo gli intervalli di confidenza per il valore di `lpsa` di due pazienti con le seguenti caratteristiche:

```
nd <- prostate[1,-9] ## to ensure correct names
nd[1,] <- c(1.45, 3.62,65,0.3,0,-0.8,7,15)
nd[2,] <- c(4,6.2,83,2.33,1,2.96,9,100)
rownames(nd) <- c("Patient A", "Patient B")
nd
```

```
          lcavol lweight age lbph svi   lcp gleason pgg45
Patient A   1.45    3.62  65 0.30   0 -0.80       7    15
Patient B   4.00    6.20  83 2.33   1  2.96       9   100
```

```
cint <- predict(fit_all, newdata = nd, interval = "confidence")
cint[,3]-cint[,2]
```

```
Patient A Patient B
 0.433730  1.782545
```

```
summary(prostate)
```

```
     lcavol           lweight           age             lbph        
 Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863  
 1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863  
 Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001  
 Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004  
 3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581  
 Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263  
      svi              lcp              gleason          pgg45       
 Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00  
 1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00  
 Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00  
 Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38  
 3rd Qu.:0.0000   3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00  
 Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00  
      lpsa        
 Min.   :-0.4308  
 1st Qu.: 1.7317  
 Median : 2.5915  
 Mean   : 2.4784  
 3rd Qu.: 3.0564  
 Max.   : 5.5829  
```

Notiamo che il paziente A è un paziente molto più "tipico" del paziente B che invece valori dei predittori abbastanza estremi se confrontanti con i valori medi delle persone nel campione: di conseguenza l'incertezza attorno alla stima per il paziente B è molto più larga.

5.
```
chosenVars <- rownames(coef(summary(fit_all))[coef(summary(fit_all))[,4] < 0.05,])
fit2 <- lm(lpsa~., data = prostate[,c(chosenVars,"lpsa")])
summary(fit2)
```

```
Call:
lm(formula = lpsa ~ ., data = prostate[, c(chosenVars, "lpsa")])

Residuals:
     Min       1Q   Median       3Q      Max 
-1.72964 -0.45764  0.02812  0.46403  1.57013 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -0.26809    0.54350  -0.493  0.62298
lcavol       0.55164    0.07467   7.388  6.3e-11 ***
lweight      0.50854    0.15017   3.386  0.00104 **
svi          0.66616    0.20978   3.176  0.00203 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7168 on 93 degrees of freedom
Multiple R-squared:  0.6264,        Adjusted R-squared:  0.6144
F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

```
cint2 <- predict(fit2, newdata = nd, interval = "confidence")
cint2[,3]-cint2[,2]
```

```
Patient A Patient B
 0.350968  1.579318
```

Gli intervalli di confidenza sono adesso meno ampi: il modello è più bilanciato perché non usa troppe variabili ed evita di andare in overfitting.

6. I due modelli sono annidati e possiamo quindi fare un test per:

$$H_0 : \beta_{age} = \beta_{lbph} = \beta_{lcp} = \beta_{gleason} = \beta_{pgg45} = 0 \; VS \text{ any } \beta_j \neq 0$$

```
anova(fit2, fit_all)
```

```
Analysis of Variance Table

Model 1: lpsa ~ lcavol + lweight + svi
Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     93 47.785
2     88 44.163  5    3.6218 1.4434 0.2167
```

Non possiamo rigettare l'ipotesi nulla: i due modelli spiegano una parte simile della varianza della variabile risposta.

Esercizio 6

Si prenda in esame il dataset `Davis` dal pacchetto R `carData`. Si desidera stimare la relazione tra il peso dichiarato (`repwt`) da uomini e donne e il peso misurato (`weight`) da uomini e donne.

1. Si stimi tre modelli di crescente complessità in cui la relazione è stimata essere la stessa per entrambi i sessi, viene permesso all'intercetta di essere diversa per i due sessi e infine in cui si permette a intercetta e coefficiente angolare di essere diversi per i due sessi.

2. Si scrivano in forma estesa (in formula matematica) i tre modelli specificati al punto 1 e si confronti la bontà di adattamento dei tre modelli indicando quale modello viene scelto come modello finale.

3. Si verifichi se per il modello selezionato valgono le assunzioni alla base della costruzione dei modelli lineari. Si commenti in particolare se sono presenti punti particolarmente influenti sulla stima.

Soluzione 6

```
data(Davis, package = "carData")
# ?carData::Davis
```

Stimiamo i tre modelli di crescente complessitá:

```
fit1 <- lm(repwt ~ weight, data = Davis)
fit2 <- lm(repwt ~ weight+sex, data = Davis)
fit3 <- lm(repwt ~ weight*sex, data = Davis)
```

Che corrispondono ai tre modelli seguenti:

$$\text{model fit1:} \quad \text{repwt}_i = \beta_0 + \beta_1 \cdot \text{weight}_i + \varepsilon_i$$

$$\text{model fit2:} \quad \text{repwt}_i = \beta_0 + \beta_1 \cdot \text{weight}_i + \beta_2 \cdot \text{male}_i + \varepsilon_i$$

$$\text{model fit2:} \quad \text{repwt}_i = \beta_0 + \beta_1 \cdot \text{weight}_i + \beta_2 \cdot \text{male}_i + \beta_3 \cdot \text{male}_i * \text{weight}_i + \varepsilon_i$$

con ε_i un termine di errore normale omoschedastico, iid a media zero, e `male` una variabile dicotomica che ha valore 1 se l'individuo i ha sesso maschile e 0 quando l'individuo i ha sesso femminile.

Possiamo confrontare i tre modelli (che sono annidati tra loro) utilizzando ANOVA (cioè facendo un test di significativitá) o attraverso dei criteri di informazione come AIC o BIC:

```
anova(fit2, fit3) # signigficant - reject H0 - interaction is useful
```

```
Analysis of Variance Table

Model 1: repwt ~ weight + sex
Model 2: repwt ~ weight * sex
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    180 7535.3
2    179 3888.3  1      3647 167.9 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit1, fit2) # signigficant - reject H0 - adding sex is useful
```

```
Analysis of Variance Table

Model 1: repwt ~ weight
Model 2: repwt ~ weight + sex
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    181 10410.2
2    180  7535.3  1    2874.9 68.675 2.599e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

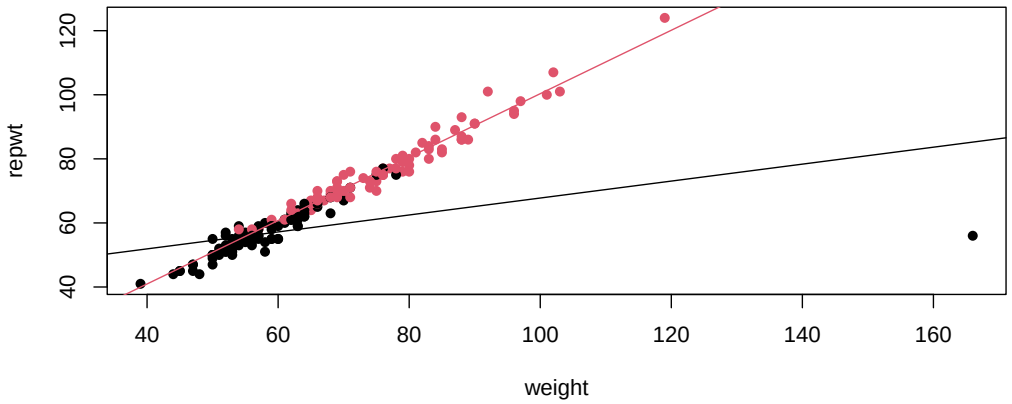
```
 AIC(fit1, fit2, fit3) # fit3 is the best model
```

```
     df      AIC
fit1  3 1264.844
fit2  4 1207.701
fit3  5 1088.621
```

```
 BIC(fit1, fit2, fit3) # fit3 is the best model
```

```
     df      BIC
fit1  3 1274.473
fit2  4 1220.539
fit3  5 1104.669
```

```
 plot(repwt ~ weight, data = Davis,
        col = ifelse(Davis$sex == "M",2,1), pch = 16)
 abline(coef(fit3)[1], coef(fit3)[2], col = 1)
 abline(coef(fit3)[1]+coef(fit3)[3], coef(fit3)[2]+coef(fit3)[4], col = 2)
```



Giá dal grafico di dispersione notiamo un punto problematico che potrebbe avere una forte influenza sulla stima

Controlliamo i grafici dei residui:

```
 par(mfrow=c(2,2))
 plot(fit3, which=c(1,2,5,4))
```

Notiamo una notevole struttura (cioè la mancanza di errore casuale) nel grafico dei fitted vs residuals: il modello ha un errore strutturale per alcune osservazioni (le donne). Il grafico del qqplot risulta anch'esso problematico e dal grafico del leverage vs residuls vediamo come una delle osservazioni risulti avere un residuo molto grande, essere un punto di leva (leverage) e avere un valore di distanza di Cook molto alto (un punto influente): l'osservazione 12 ha una grande influenza sulla stima del modello. Cosa succede se la rimuoviamo?

```
fit1_minus12 <- lm(repwt ~ weight, data = Davis, subset = -12)
fit2_minus12 <- lm(repwt ~ weight+sex, data = Davis, subset = -12)
fit3_minus12 <- lm(repwt ~ weight*sex, data = Davis, subset = -12)

AIC(fit1_minus12, fit2_minus12, fit3_minus12)
```

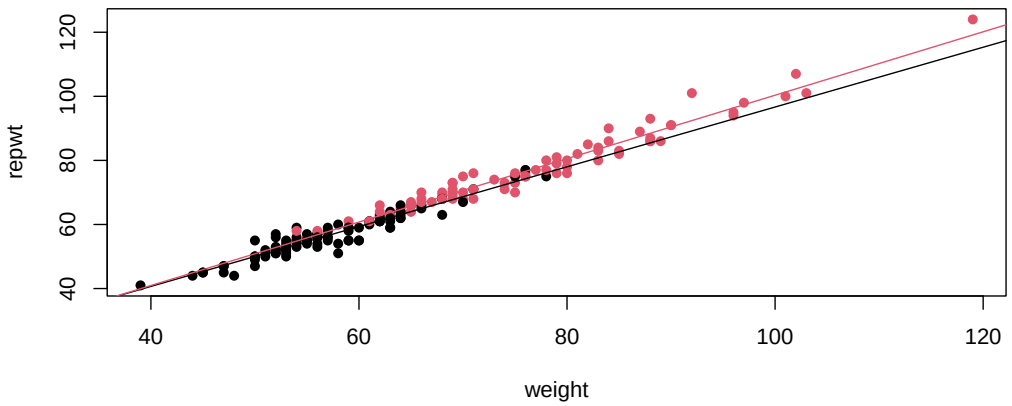
```
             df      AIC
fit1_minus12  3 826.5627
fit2_minus12  4 817.6706
fit3_minus12  5 817.4721
```

```
# fit3 still better but less spectacular difference
BIC(fit1_minus12, fit2_minus12, fit3_minus12)
```

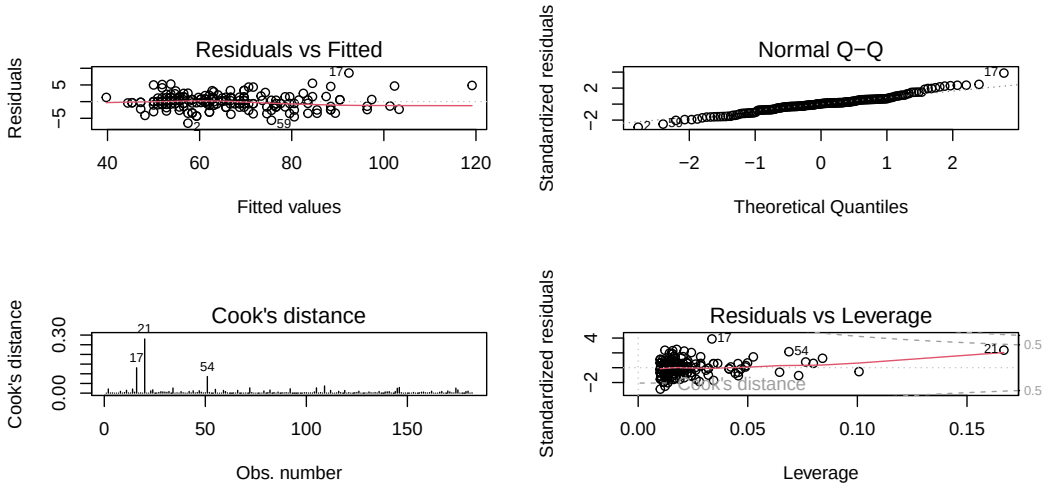
```
             df      BIC
fit1_minus12  3 836.1747
fit2_minus12  4 830.4866
fit3_minus12  5 833.4921
```

```
# fit2 better

plot(repwt ~ weight, data = Davis, subset = -12,
       col = ifelse(Davis$sex == "M",2,1), pch = 16)
abline(coef(fit3_minus12)[1], coef(fit3_minus12)[2], col = 1)
abline(coef(fit3_minus12)[1]+coef(fit3_minus12)[3],
         coef(fit3_minus12)[2]+coef(fit3_minus12)[4], col = 2)
```




```
par(mfrow=c(2,2))
plot(fit3_minus12, which=c(1,2,5,4))
```



Le assunzioni alla base del modello sono ora più credibili e non ci sono punti eccessivamente influenti sulla stima del modello.

Esercizio 7

[Esercizio di Esame aa 2019/2020 - prof. Gaetan]

Si consideri il modello di regressione

$$Y_i = \begin{cases} \beta_1 + \varepsilon_i & i = 1, \dots, 5 \\ \beta_1 + \beta_2(i-5) + \varepsilon_i & i = 6, \dots, 10 \end{cases}$$

dove ε_i, $i = 1, \dots, 10$ sono v.c. casuali $\mathcal{N}(0, \sigma^2)$ indipendenti.

1. Si specifichi se le assunzioni usualmente adottate in un modello di regressione lineare Gaussiano (linearità della relazione, normalità ed omoschedasticità degli errori, indipendenza delle osservazioni) sono soddisfatte dal modello sopra riportato.

2. Il modello può essere scritto nella forma matriciale $Y = X\beta + \varepsilon$. Si dia l'espressione della matrice X.

3. Si argomenti quale possa essere la distribuzione dello stimatore di massima verosimiglianza per β.

4. Supponendo di aver ottenuto le stime $(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}) = (2.86, 0.1, 0.86)$, si derivino gli intervalli di confidenza con livello esatto 0.90 per β_1 e β_2.

5. Di quale altra informazione ci sarebbe bisogno per calcolare R^2 ?

Soluzione 7

1. Si, le assunzioni sono rispettate

2.

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 6 \\ 1 & 7 \\ 1 & 8 \\ 1 & 9 \\ 1 & 10 \end{bmatrix}$$

3. Dato che gli errori sono distribuiti secondo una normale si ha che $Y|X = x$ segue una normale e la stima di massima verosimiglianza di β corrisponde a $\hat{\beta} = (X^\top X)^{-1} Xy$, una combinazione lineare delle osservazioni, che sono un campione estratto da una normale di standard, per cui anche $\hat{\beta}$ segue una normale standard (questo risultato è esatto e non basato sul fatto che tutti gli stimatori di massima verosimiglianza sono approssimativamente normante distribuiti per $n \to \infty$)

4. Dobbiamo costruire la matrice di varianza-covarianza stimata per gli stimatori che sappiamo avere la forma di:

   $$Var(\hat{\beta}) = \hat{\sigma}(X^\top X)$$

   In prima istanza dobbiamo quindi derivare la matrice $(X^\top X)$

   $$(X^\top X) = \begin{bmatrix} 10 & 40 \\ 40 & 330 \end{bmatrix}$$

   da cui deriviamo che

   $$(X^\top X)^{-1} = \begin{bmatrix} 0.190 & 0.0240 \\ 0.024 & 0.0059 \end{bmatrix}$$

   e

   $$Var(\beta) = \begin{bmatrix} 0.1450 & -0.01750 \\ -0..01750 & 0 - 00438 \end{bmatrix}$$

   Per cui:

```
# ci for beta_1
c(2.86 + qt(.05, df = 8)*sqrt(0.1450), 2.86 + qt(.95, df = 8)*sqrt(0.1450))
```

```
[1] 2.151905 3.568095
```

```
# ci for beta_2
c(0.1 + qt(.05, df = 8)*sqrt(0.00438), 0.1 + qt(.95, df = 8)*sqrt(0.00438))
```

```
[1] -0.02306781  0.22306781
```

In alternativa con R si poteva derivare il tutto con:

```
X <- cbind(rep(1,10),c(rep(0,5), seq(6,10)))
se_betas <- 0.86*sqrt(diag(solve(t(X) %*% X)))
# ci for beta_1
c(2.86 + qt(.05, df = 8)*se_betas[1], 2.86 + qt(.95, df = 8)*se_betas[1])
```

```
[1] 2.155407 3.564593
```

```
# ci for beta_2
c(0.1 + qt(.05, df = 8)*se_betas[2], 0.1 + qt(.95, df = 8)*se_betas[2])
```

```
[1] -0.02265391  0.22265391
```

5. Per derivare il valore di R^2 dovremmo poter essere in grado di calcolare $\sum(y_i - \overline{y})^2$: manca l'informazione sulle y_i